

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه تهران

دفتر آموزش های حرفه ای و تخصصی

عنوان:

پاک سازی داده دیابت

پروژه دیابت سرخ پوستان پیما

استاد:

مهندس محمدرضا محتاط

نام دانشجو:

زهرا غلامی مندی

تیرماه ۱۴۰۳

چکیده

در این مستند گام‌های مربوط به فازهای متدولوژی CRISP-DM در پروژه دیابت سرخ‌پوستان پیمما مورد بررسی قرار داده شده است. دیتاست دیابت سرخ‌پوستان پیمما یکی از مجموعه داده‌های شناخته شده در زمینه پیش‌بینی دیابت نوع ۲ است که توسط مؤسسه ملی دیابت و بیماری‌های گوارشی و کلیه^۱ تهیه شده است. این دیتاست شامل ۷۶۸ نمونه و ۸ ویژگی ورودی به همراه یک ویژگی خروجی است که نشان می‌دهد آیا فرد مبتلا به دیابت است یا خیر. این دیتاست به دلیل چالش‌های موجود در پیش‌بینی دقیق دیابت و اهمیت ویژگی‌های ورودی در تشخیص بیماری، به طور گسترده در تحقیقات و پروژه‌های یادگیری ماشین استفاده می‌شود. من در این پروژه به بررسی و ارزیابی و بهبود تکنیک‌های پیش‌پردازش داده‌ها، مانند مدیریت داده‌های گمشده و نامتوازن می‌پردازم.

فهرست مطالب

صفحه	عنوان
۱	فصل ۱ درک کسب و کار و داده
۱-۱	مقدمه
۱-۲	شناسایی داده ها
۱-۲-۱	ستون اول؛ Pregnancies
۱-۲-۲	ستون دوم؛ Glucose
۱-۲-۳	ستون سوم؛ Blood Pressure
۱-۲-۴	ستون چهارم؛ Skin Thickness
۱-۲-۵	ستون پنجم؛ Insulin
۱-۲-۶	ستون ششم؛ BMI
۱-۲-۷	ستون هفتم؛ Diabetes Pedigree Function
۱-۲-۸	ستون هشتم؛ Age
۱-۲-۹	ستون نهم؛ Outcome
۱-۳	تشخیص خطا یا نویز
۱-۴	ساختار کلی پروژه
۷	فصل ۲ آماده سازی داده
۲-۱	مقدمه
۲-۲	وارد کردن داده
۲-۳	ساختار بندی و مجتمع کردن داده
۲-۴	رفع خطا یا نویز
۲-۵	شناسایی داده های پرت
۲-۵-۱	تشخیص توزیع داده ها

۱۳	۲-۵-۲. شناسایی تعداد داده های پرت
۱۵	۲-۵-۳. استفاده از روش ۱.۵ تا ۳ IQR
۱۶	۲-۵-۴. استفاده از روش ۳ تا ۵ سیگما
۱۷	۲-۶. شناسایی داده های مفقوده
۱۷	۲-۶-۱. تشخیص توزیع داده ها
۱۷	۲-۶-۲. شناسایی درصد داده های مفقوده
۱۸	۲-۶-۳. استفاده از روش Random-Normal
۱۹	۲-۷. هم مقیاس سازی داده ها

Error! Bookmark not defined.

نتیجه گیری

فصل ۱

درک کسب و کار و داده

۱-۱. مقدمه

در فصل اول این پروژه، به بررسی و انجام فاز اول متدولوژی CRISP-DM خواهیم پرداخت. این فاز که به عنوان "درک کسب و کار و داده" شناخته می شود، شامل مراحل کلیدی برای ایجاد یک پایه قوی جهت تحلیل و مدل سازی داده ها است. در این مرحله، ابتدا به شناسایی و تعریف اهداف کسب و کار و تحقیق پرداخته و سپس سؤالات کلیدی و نیازمندی های پروژه مشخص می شوند. بعد از آن، به جمع آوری و بررسی اولیه داده ها خواهیم پرداخت تا با ساختار و ویژگی های آن ها آشنا شویم. تحلیل اولیه داده ها به ما کمک می کند تا نواقص و مشکلات موجود در داده ها را شناسایی کرده و استراتژی های مناسب برای پیش پردازش داده ها را برنامه ریزی کنیم. در نهایت، نتایج این فاز به ما کمک خواهد کرد تا مسیر درست برای مراحل بعدی پروژه را تعیین کنیم.

۱-۲. شناسایی داده‌ها

شناسایی داده‌ها یکی از مراحل ابتدایی و حیاتی در فاز اول متدولوژی CRISP-DM است. در این مرحله، به جمع‌آوری و بررسی اولیه داده‌ها می‌پردازیم تا ساختار، کیفیت و ویژگی‌های کلیدی آن‌ها را درک کنیم. این شناخت اولیه به ما کمک می‌کند تا مشکلات احتمالی را شناسایی کرده و استراتژی‌های مناسب برای پردازش و تحلیل داده‌ها را تعیین کنیم.

۱-۲-۱. ستون اول؛ Pregnancies

این ویژگی نشان‌دهنده تعداد بارداری‌های قبلی هر زن در دیتاست است. تعداد بارداری‌ها می‌تواند به‌عنوان یک عامل خطر برای دیابت نوع ۲ باشد، زیرا تغییرات هورمونی و فیزیولوژیکی در دوران بارداری ممکن است بر سیستم متابولیکی تاثیر بگذارد.

۱-۲-۲. ستون دوم؛ Glucose

این ویژگی نشان‌دهنده غلظت گلوکز پلاسما در دو ساعت پس از انجام آزمایش تحمل گلوکز خوراکی (mg/dl) است. گلوکز بالا می‌تواند نشان‌دهنده مشکلات متابولیسم قند و احتمال دیابت باشد. این شاخص یکی از مهم‌ترین ویژگی‌ها برای تشخیص دیابت است.

۱-۲-۳. ستون سوم؛ Blood Pressure

این ویژگی نشان‌دهنده فشارخون دیاستولیک (mm Hg) است. فشارخون بالا ممکن است با افزایش خطر دیابت و بیماری‌های قلبی مرتبط باشد. بررسی فشارخون به‌عنوان یک ویژگی

کلیدی در ارزیابی سلامت قلب و عروق اهمیت دارد.

۱-۲-۴. ستون چهارم؛ Skin Thickness

این ویژگی نشان‌دهنده ضخامت پوست در محل سه سر (mm) است. ضخامت پوست می‌تواند با چاقی و مقاومت به انسولین مرتبط باشد. این ویژگی به‌عنوان یک نشانگر برای ارزیابی وضعیت چربی بدن و ذخایر انرژی بدن اهمیت دارد.

۱-۲-۵. ستون پنجم؛ Insulin

این ویژگی نشان‌دهنده سطح انسولین سرم در دو ساعت پس از آزمایش ($\mu\text{U/ml}$) است. سطح انسولین می‌تواند نشان‌دهنده عملکرد پانکراس و مقاومت به انسولین باشد. این ویژگی می‌تواند به شناسایی مشکلات در تولید یا استفاده از انسولین کمک کند.

۱-۲-۶. ستون ششم؛ BMI

شاخص توده بدنی (BMI) به‌صورت وزن تقسیم بر قد محاسبه می‌شود. BMI بالا می‌تواند نشانه‌ای از چاقی و خطر بالای دیابت باشد. این شاخص یکی از معیارهای مهم برای ارزیابی وضعیت بدنی و چاقی است.

۱-۲-۷. ستون هفتم؛ Diabetes Pedigree Function

این ویژگی نشان‌دهنده احتمال ارثی دیابت است و به‌ارث‌بردن دیابت در خانواده را نشان

می‌دهد. این فاکتور نشان‌دهنده ریسک ارثی برای دیابت است و می‌تواند به تشخیص افرادی که به طور ژنتیکی مستعد دیابت هستند کمک کند.

۸-۲-۱. ستون هشتم؛ Age

این ویژگی سن فرد را نشان می‌دهد. سن بالاتر می‌تواند با افزایش خطر دیابت مرتبط باشد. بررسی سن به‌عنوان یک فاکتور مهم در ارزیابی ریسک بیماری‌های مختلف، از جمله دیابت، اهمیت دارد.

۹-۲-۱. ستون نهم؛ Outcome

این ویژگی خروجی که ستون هدف نامیده می‌شود، نشان‌دهنده وضعیت دیابت فرد است (۰ برای عدم ابتلا و ۱ برای ابتلا به دیابت). این فیلد برچسب هدف است و نشان می‌دهد آیا فرد دیابت دارد یا خیر.

۳-۱. تشخیص خطا یا نویز

تشخیص و تصحیح خطا و نویز در داده‌ها یکی از مراحل حیاتی در پیش‌پردازش داده‌ها است. مقادیر غیرمنطقی یا خارج از محدوده‌های مجاز باید شناسایی و تصحیح شوند تا از تأثیر منفی آن‌ها بر مدل‌های پیش‌بینی جلوگیری شود.

شکل ۱-۱. مقادیری که ویژگی‌ها نمی‌پذیرند.

Blood Pressure	Glucose	Pregnancies
صفر و منفی	منفی و صفر	منفی و بازه بیشتر از ۱۷
BMI	Insulin	Skin Thickness
کمتر از ۱۰	صفر	صفر
Outcome	Age	Diabetes Pedigree Function
مقادیر متفاوت از اعداد ۰ و ۱	کمتر از ۲۱	کمتر از صفر و بیشتر از ۲.۵

۱-۴. ساختار کلی پروژه

در فصل اول این پروژه، به بررسی و اجرای فاز اول متدولوژی CRISP-DM یعنی درک کسب‌وکار و داده، پرداختیم. ابتدا با تعریف اهداف کسب‌وکار و تحقیق، سؤالات کلیدی و نیازمندی‌های پروژه را مشخص کردیم. سپس به شناسایی و بررسی ویژگی‌های مختلف دیابت سرخ‌پوستان پیما پرداخته و هر ویژگی را به طور جداگانه توضیح داده شد. اطلاعات به‌دست‌آمده در این مرحله، پایه‌ای قوی برای مراحل بعدی پروژه فراهم می‌کند.

فصل ۲

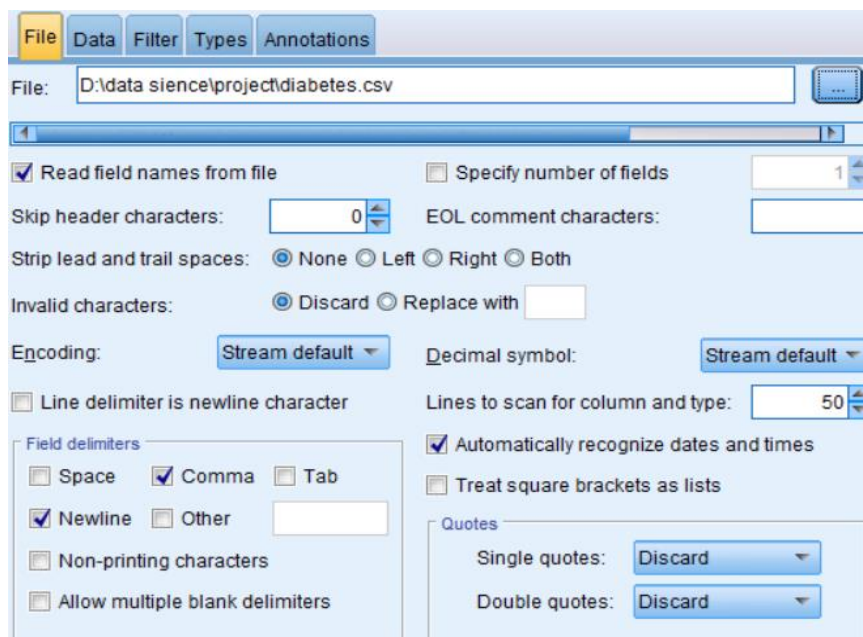
آماده‌سازی داده

۲-۱. مقدمه

در فصل دوم این پروژه، به مرحله آماده‌سازی داده‌ها می‌پردازیم. آماده‌سازی داده‌ها شامل مجموعه‌ای از فعالیت‌هاست که به منظور بهبود کیفیت داده‌ها، افزایش دقت مدل‌ها و کاهش نویز و خطاهای موجود انجام می‌شود. در این فصل تکنیک‌های مختلف پیش‌پردازش داده‌ها را به کار خواهیم گرفت تا داده‌ها را برای تحلیل و مدل‌سازی آماده کنیم. ابتدا داده‌ها را وارد و سپس به پاک‌سازی داده‌ها خواهیم پرداخت که شامل شناسایی و حذف داده‌های گمشده و نویزهای موجود در دیتاست می‌شود. سپس به نرمال‌سازی و استانداردسازی ویژگی‌ها می‌پردازیم تا مقادیر هر ویژگی در محدوده مناسب قرار گیرد.

۲-۲. وارد کردن داده

ابتدا دیتاست مربوط به دیابت سرخ‌پوستان پیما را از سایت Kaggle دریافت می‌کنیم که یک فایل با پسوند csv به ما می‌دهد که شامل ۹ ویژگی و ۷۶۸ رکورد است. سپس باید این فایل csv را در IBM وارد کنیم، با استفاده از تب Source، ابزار Var.File را به محیط کار اضافه کرده و به فایل Csv متصل می‌کنیم.



شکل ۱-۲. Import کردن داده‌ها در IBM

۱۰ رکورد اول بصورت زیر نمایش داده می‌شوند:

Preview from diabetes.csv Node (9 fields, 10 records)

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
1	6	148	72	35	0	33....	0.627	50
2	1	85	66	29	0	26....	0.351	31
3	8	183	64	0	0	23....	0.672	32
4	1	89	66	23	94	28....	0.167	21
5	0	137	40	35	168	43....	2.288	33
6	5	116	74	0	0	25....	0.201	30
7	3	78	50	32	88	31....	0.248	26
8	10	115	0	0	0	35....	0.134	29
9	2	197	70	45	543	30....	0.158	53
10	8	125	96	0	0	0.0...	0.232	54

شکل ۲-۲. ۱۰ رکورد اول دیتاست

مرحله بعد پس از وارد کردن داده، مشخص کردن Data Type داده‌ها است:

جدول ۲-۱. نوع داده بر اساس تعریف آنها

Blood Pressure	Diabetes Pedigree Function		Glucose	Pregnancies
Continuous	Continuous		Continuous	Continuous
Skin Thickness	Outcome	Insulin	BMI	Age
Continuous	Flag	Continuous	Continuous	Continuous

اما نوع داده‌ای که نرم‌افزار IBM به ما نشان می‌دهد:

Field	Measurement	Values	Missing	Check	Role
Pregnancies	Continuous	[0, 17]		None	Input
Glucose	Continuous	[0, 199]		None	Input
BloodPressu...	Continuous	[0, 122]		None	Input
SkinThickness	Continuous	[0, 99]		None	Input
Insulin	Continuous	[0, 846]		None	Input
BMI	Continuous	[0.0, 67.1]		None	Input
DiabetesPed...	Continuous	[0.078, 2.42]		None	Input
Age	Continuous	[21, 81]		None	Input
Outcome	Flag	1/0		None	Target

شکل ۲-۳. نوع داده وارد شده در IBM

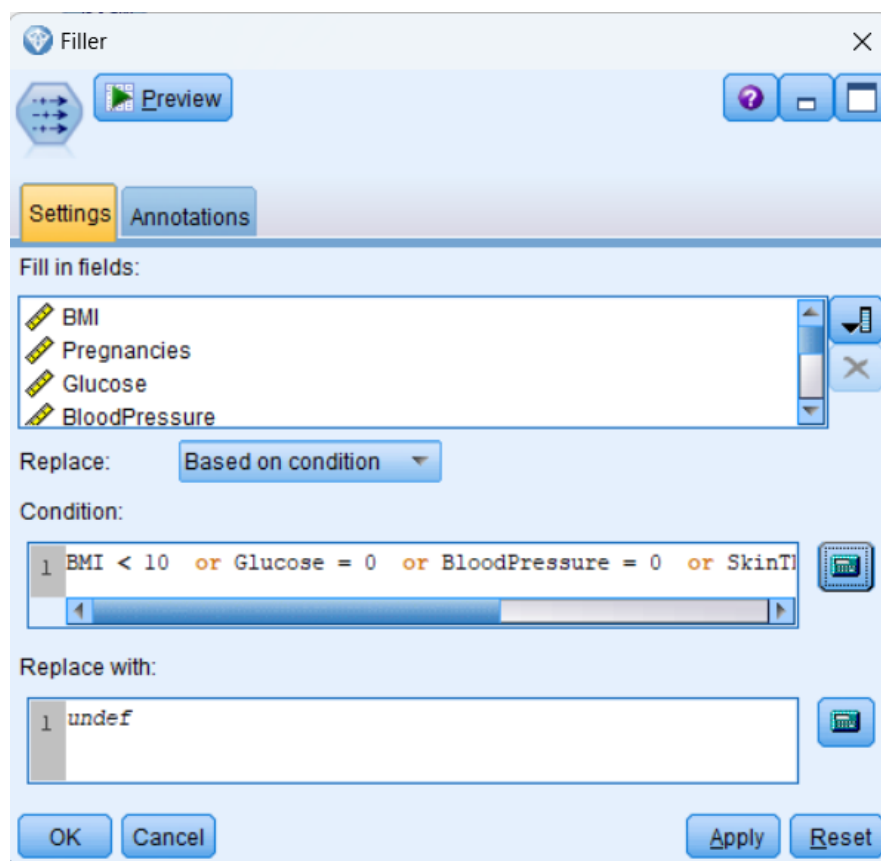
۲-۳. ساختاربندی و مجتمع کردن داده

در این پروژه نیازی به ساختاربندی و مجتمع کردن داده نداریم؛ زیرا داده‌ها پراکنده نیستند و در یک دیتاست ذخیره شده‌اند.

۲-۴. رفع خطایا نویز

در این مرحله با کمک جدول ۱-۱ و نوع توزیع می‌توانیم به شناسایی و بررسی نویز و خطا بپردازیم و سپس با کمک ابزار Filler نویز یا خطا را مشخص و سپس در مراحل بعد آن را رفع می‌کنیم.

شکل ۲-۴. مشخص کردن نویز و خطا با استفاده از ابزار Filler



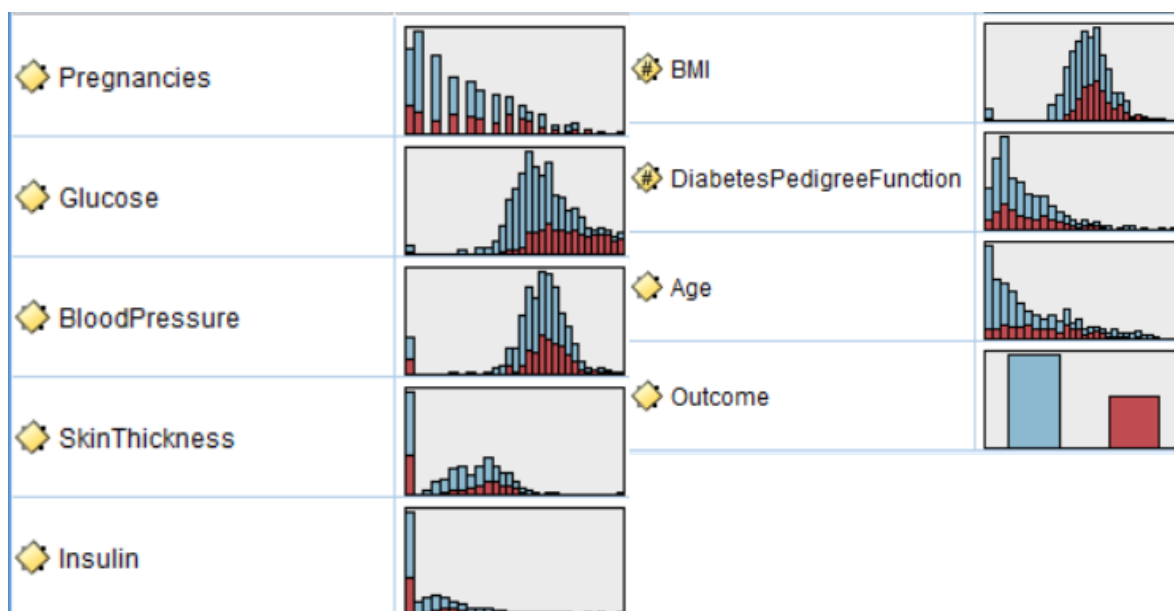
۲-۵. شناسایی داده‌های پرت

۲-۵-۱. تشخیص توزیع داده‌ها

اولین گام اجرایی در شناسایی داده‌های پرت، مشخص نمودن توزیع داده‌ها است؛ بنابراین در این مرحله دوراه برای شناسایی توزیع داده‌ها وجود دارد:

۲-۵-۱-۱. روش اول؛ شناسایی با کمک نمودار

در این روش با استفاده از ابزار Data Audit در تب Output به صورت چشمی با استفاده از نمودار می‌توانیم توزیع داده‌ها را مشاهده کنیم.



شکل ۲-۵. نمایش توزیع داده‌ها با کمک Data Audit

با کمک این روش می‌تواند تشخیص داد که ویژگی‌های BMI، BloodPressure، Glucose دارای توزیع نرمال هستند.

۲-۵-۱. روش دوم؛ با استفاده از ابزار Sim Fit

برای تشخیص توزیع ویژگی‌ها در این روش از ابزار Sim Fit در تب Output و با استفاده از دو آزمون کولموگروف-اسمیرنوف و اندرسون دارلینگ متوجه نوع توزیع هر یک از ویژگی‌ها می‌شویم:

Field	Storage	Status		Distribution
Pregnancies	Integer		<input type="checkbox"/>	Exponential
Glucose	Integer		<input type="checkbox"/>	Normal
BloodPressure	Integer		<input type="checkbox"/>	Normal
SkinThickness	Integer		<input type="checkbox"/>	Exponential
Insulin	Integer		<input type="checkbox"/>	Exponential
BMI	Real		<input type="checkbox"/>	Normal
DiabetesPedigr...	Real		<input type="checkbox"/>	Lognormal
Age	Integer		<input type="checkbox"/>	Lognormal
Outcome	Integer		<input type="checkbox"/>	Categorical

شکل ۲-۶. تشخیص توزیع ویژگی‌ها با استفاده از ابزار Sim Fit

با استفاده از این دو روش متوجه می‌شویم که ویژگی‌های BMI، BloodPressure، Glucose دارای توزیع نرمال هستند.

۲-۵-۲. شناسایی تعداد داده‌های پرت

میتوان با استفاده از Data Audit و سپس در تب Quality با تنظیم روش ۳ تا ۵ سیگما یا

۱.۵ تا ۳ IQR (بسته به توزیع داده) داده‌های پرت و خیلی پرت را برای هر مؤلفه مشاهده کرد.

Field	Measurement	Outliers	Extremes
Pregnancies	Continuous	4	0
Glucose	Continuous	5	0
BloodPressu...	Continuous	35	0
SkinThickness	Continuous	1	0
Insulin	Continuous	15	3
BMI	Continuous	14	0
DiabetesPed...	Continuous	7	4
Age	Continuous	5	0
Outcome	Flag	--	--

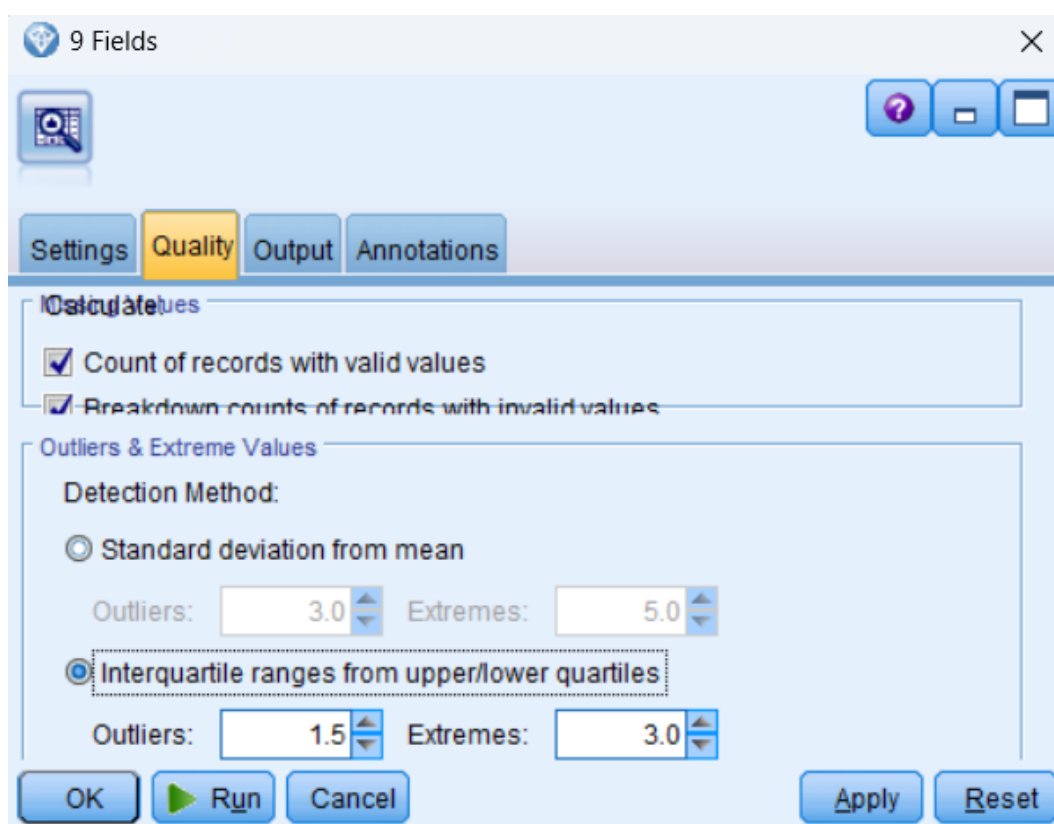
شکل ۲-۷. داده‌های پرت و خیلی پرت با استفاده از روش ۳ سیگما

Field	Measurement	Outliers	Extremes
Pregnancies	Continuous	4	0
Glucose	Continuous	5	0
BloodPressu...	Continuous	10	35
SkinThickness	Continuous	1	0
Insulin	Continuous	26	8
BMI	Continuous	18	1
DiabetesPed...	Continuous	23	6
Age	Continuous	9	0
Outcome	Flag	--	--

شکل ۲-۸. داده‌های پرت و خیلی پرت با روش ۱.۵ تا ۳ IQR

۳-۵-۲. استفاده از روش ۱.۵ تا ۳ IQR

برای مؤلفه‌های Insulin, Skin Thickness, Pregnancies، با استفاده از Data Audit و با انجام تنظیمات زیر، داده‌های پرت و خیلی پرت را از روش ۱.۵ تا ۳ IQR مدیریت می‌کنیم:



شکل ۲-۹. تنظیم ۱.۵ تا ۳ IQR در Data Audit

سپس برای مؤلفه‌های فوق‌الذکر از طریق Action‌هایی که در شکل زیر نمایش داده شده است سوپر نود Outlier & Extreme را ایجاد می‌کنیم:

Field	Measurement	Outliers	Extremes	Action
Pregnancies	Continuous	4	0	Coerce
Glucose	Continuous	5	0	None
BloodPressu...	Continuous	10	35	None
SkinThickness	Continuous	1	0	Coerce
Insulin	Continuous	26	8	Coerce
BMI	Continuous	18	1	None
DiabetesPed...	Continuous	23	6	None
Age	Continuous	9	0	None
Outcome	Flag	--	--	--

شکل ۲-۱۰. مدیریت داده‌های پرت و خیلی پرت مؤلفه‌های غیرنرمال

۴-۵-۲. استفاده از روش ۳ تا ۵ سیگما

حال برای مؤلفه‌هایی که دارای توزیع نرمال و لاگ نرمال هستند از این روش استفاده

می‌کنیم.

Field	Measurement	Outliers	Extremes	Action
Pregnancies	Continuous	0	0	None
Glucose	Continuous	5	0	Nullify
BloodPressu...	Continuous	35	0	Coerce
SkinThickness	Continuous	1	0	None
Insulin	Continuous	0	0	None
BMI	Continuous	14	0	Coerce
DiabetesPed...	Continuous	7	4	Coerce outliers / nullify extremes
Age	Continuous	5	0	Coerce
Outcome	Flag	--	--	--





























شکل ۲-۱۱. مدیریت داده‌های پرت و خیلی پرت مؤلفه‌های نرمال و لاگ نرمال

حالا کار شناسایی داده‌های پرت به پایان می‌رسد.

۲-۶. شناسایی داده‌های مفقوده

۲-۶-۱. تشخیص توزیع داده‌ها

از آنجایی که قبل از عمل شناسایی داده‌های پرت ما توزیع داده‌ها را مشاهده کردیم اکنون بعد عمل شناسایی داده‌های پرت و مدیریت آنها بار دیگر باید به وسیله Sim Fit توزیع داده‌ها را مشاهده کنیم.

Field	Storage	Status		Distribution
Pregnancies	 Real			Exponential
Glucose	 Integer			Lognormal
BloodPressure	 Real			Weibull
SkinThickness	 Real			Normal
Insulin	 Real			Normal
BMI	 Real			Normal
DiabetesPedigree...	 Real			Lognormal
Age	 Real			Lognormal
Outcome	 Integer			Categorical

شکل ۲-۱۲. نمایش مجدد توزیع داده‌ها با Sim Fit

۲-۶-۲. شناسایی درصد داده‌های مفقوده

ابتدا با استفاده از Data Audit، درصد داده‌های مفقوده هر فیلد را بررسی می‌کنیم.

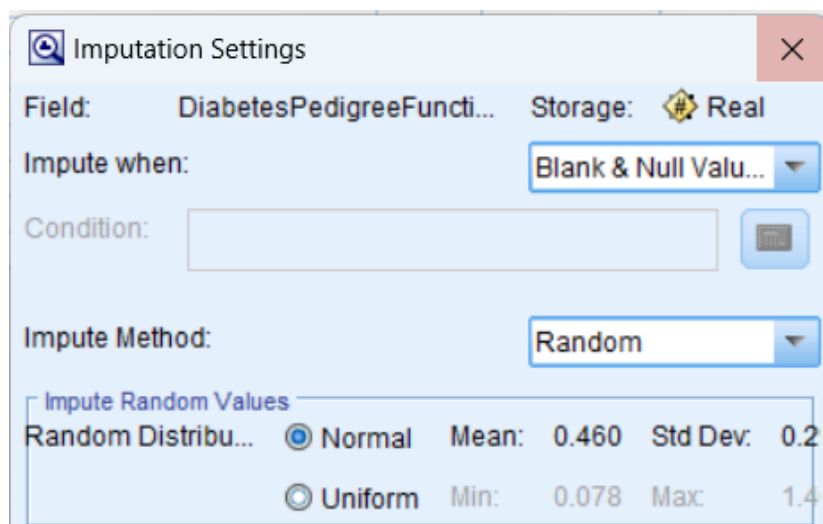
Field	Measurement	Outliers	Extremes	% Complete	Valid Records	Null Value
# Pregnancies	Continuous	0	0	100	768	0
# Glucose	Continuous	0	0	100	768	0
# BloodPressu...	Continuous	0	0	100	768	0
# SkinThickness	Continuous	0	0	99.87	767	1
# Insulin	Continuous	0	0	98.958	760	8
# BMI	Continuous	0	0	100	768	0
# DiabetesPed...	Continuous	0	0	100	768	0
# Age	Continuous	0	0	100	768	0
# Outcome	Flag	--	--	100	768	0

شکل ۲-۱۳. درصد داده‌های مفقوده مؤلفه‌ها

واضح است که ما برای ویژگی‌های Diabetes Pedigree Function و Glucose داده مفقوده داریم؛ بنابراین می‌بایست تنها برای این ویژگی‌ها مدیریت داده مفقوده را انجام دهیم.

۲-۶-۳. استفاده از روش Random-Normal

برای ویژگی‌هایی که نرمال یا لاگ نرمال هستند؛ یعنی مؤلفه‌های Diabetes Pedigree Function، Glucose، BloodPressure، BMI، Age از این روش استفاده می‌کنیم.



در نهایت سوپر نود Missing value imputation را ایجاد می‌کنیم و اگر با Data Audit مشاهده کنیم دیگر داده مفقوده‌ای نداریم و تمام فیلدها ۱۰۰ درصد پر شده‌اند.

Field	Measurement	Outliers	Extremes	% Complete	Valid Records	Null Value
# Pregnancies	Continuous	0	0	100	768	0
# Glucose	Continuous	0	0	100	768	0
# BloodPressu...	Continuous	0	0	100	768	0
# SkinThickness	Continuous	0	0	100	768	0
# Insulin	Continuous	0	0	100	768	0
# BMI	Continuous	0	0	100	768	0
# DiabetesPed...	Continuous	0	0	100	768	0
# Age	Continuous	0	0	100	768	0
# Outcome	Flag	--	--	100	768	0

شکل ۱۴-۲. نمایش مدیریت داده‌های پرت و مفقوده کل مولفه‌ها

در اینجا کار مدیریت و شناسایی داده‌های مفقوده به پایان می‌رسد. طی این فرایند، داده‌های مفقوده شناسایی و با استفاده از روش‌های مناسب جایگزین شدند تا از تأثیر منفی آن‌ها بر تحلیل‌ها و مدل‌سازی‌های آینده جلوگیری شود.

۷-۲. هم‌مقیاس‌سازی داده‌ها

در این بخش به مقیاس‌سازی داده‌ها پرداخته می‌شود روش‌های مختلفی برای مقیاس‌سازی وجود دارد، از جمله نرمال‌سازی و استانداردسازی که در این پروژه از روش نرمال‌سازی برای هم‌مقیاس‌سازی داده‌ها استفاده کرده‌ام؛ زیرا نرمال‌سازی، داده‌ها را به یک محدوده مشخص مانند [0,100] تبدیل می‌کند.

Transform Continuous Field

☒ Put all continuous input fields on a common scale (highly recommended if feature construction will be performed)

Rescaling method: **Min/Max transformation** Minimum: **0.0** Maximum: **100.0**

☐ Rescale a continuous target with a Box-Cox transformation to reduce skew

Final mean: **0.0** Final standard deviation: **1.0**

شکل ۲-۱۵. هم‌مقیاس‌سازی با روش Min/Max

و در آخر نیاز است که از داده‌های تمیز شده خروجی اکسل بگیریم، برای این کار کافیست ابزار Excel را از تب Export بر روی صفحه کار قرار داده:

Export Publish Annotations

File name: **D:\data science\diabetes-project\Z-GholamiMendi-diabetes-Proje**

File type: **Excel 2007-2013 (*.xlsx)**

Options

☒ Create new file ☐ Insert into existing file

☒ Include field names Start in cell: **A1**

Choose worksheet: ☒ By index **0** ☐ By name **Sheet1**

☐ Launch Excel(tm)

☐ Generate an import node for this data

شکل ۲-۱۶. گرفتن خروجی برای نرم افزار Excel

نتیجه‌گیری

در این فصل به آماده‌سازی داده‌ها که شامل مدیریت نویز، شناسایی داده‌های پرت، مدیریت داده‌های مفقوده و هم‌مقیاس‌سازی داده‌ها است پرداختیم، تشخیص ناهنجاری و مدیریت داده‌های نامتوازن را می‌بایست؛ ولی به علت اینکه آموزش داده نشده است به زمان دیگر موکول شد.

